Title	複雑ネットワーク分析ツールを用いた日本語文章の視覚化 (数学ソフトウェアと教育 : 数学ソフトウェアの効果的利用に関する研究)
Author(s)	吉澤, 康介; 三宅, 修平
Citation	数理解析研究所講究録 (2012), 1780: 111-118
Issue Date	2012-03
URL	http://hdl.handle.net/2433/171826
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
第1780巻 2012年 111-118

複雑ネットワーク分析ツールを用いた日本語文章の視覚化

東京情報大学 総合情報学部　吉澤 康介 (Kousuke YOSHIZAWA)
三宅 修平 (Shuhei MIYAKE)
Faculty of Informatics,
Tokyo University of Information Sciences

1 はじめに

複雑な概念を理解するために，我々は対象を視覚化することがある．視覚化の具体例として，各種のチャートや表，箇条書き，丸と矢印，などが挙げられる．

このうち，「丸と矢印」を数学的に整理したものが，グラフである．

我々の持つ知識や概念をグラフ，すなわち，ネットワーク構造として表現する試みは，古くから行われてきている．かつては，「紙と鉛筆」で行っていた視覚化の作業が，情報技術の進展に伴って，大量のデータを機械的に処理して視覚化できるようになってきている．その中でも，本論文で着目しているのは，複雑ネットワーク [2, 3, 4] という概念と，その分析・視覚化のために開発された各種ツール類である．

複雑ネットワークとは，現実世界の巨大なネットワークの性質について研究する手法である．現実世界には，多様なネットワークが存在する．例えば，友人関係，Web のリンク構造，論文の参照関係などである．興味深いことに，これらの全く異なるネットワークに，ある一定の共通の性質を見出すことができる．その代表的な性質は，「スケールフリー性」，「スモールワールド性」，「クラスター性」などである．

近年，この複雑ネットワークに関して様々な知見が得られており，また，複雑ネットワークの研究用として，いくつかの視覚化ツールの開発・改良が続いている．代表的なツールとしては，pajek [9] や Cytoscape [8] などがある．

本論文の基本的なアイデアは，こういった複雑ネットワークの知見やツールを利用して，複雑な概念を (可能な限り) 機械的に視覚化し，概念の理解を手助けすることが可能かどうか，検証を試みるという点にある．

2 日本国憲法の視覚化の試み

2.1 視覚化の基本的な手順

本論文では，具体的には，次のようなことを試みた．

1. まず，「複雑な概念」の例として，法律の条文 (日本国憲法) を取り上げる．法律の条文を選択したのは，
   - 文の内容が，一般的な文章に比べて論理的に整理されていると期待できる

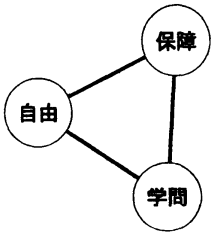

図 1: 日本国憲法第二十三条のネットワーク

- その一方で，法律の専門家でない者にとって，内容がなかなか理解しづらい

という理由による．

2. ひとつの条文の中に出現する名詞には「相互に関係がある」と考え，名詞をノードとするネットワークを考える．例えば,「第二十三条 学問の自由は，これを保障する.」という条文の場合であれば，図 1 に示すようなネットワークを考える．

   なお，単語 (名詞) の抽出には，形態素解析ツール Chasen [12] を使用している．

3. これを法律全体に対して実施する．当然，同じ単語が複数の条文に出現するので，(少数の例外を除いて) 法律全体が一つの連結したネットワークとなる．

4. このネットワークを視覚化する．なお，本論文では，視覚化ツールとして Cytoscape を使用している．

2.2 視覚化の第一段階

図 2 は，以上のようにして作成した，日本国憲法全体のネットワークである．このネットワークは，597 個のノードと，14578 のエッジからなる．

データとして使用したのは，第一条から第百三条までの条文の本文部分である．前文と章名，条名は利用していない．

Cytoscape に搭載されているレイアウトアリゴリズムのひとつである Spring-Embedded Layout(ノード同士が相互に反発しあい，エッジがバネでてきているとした状態で，全体のエネルギーを最小化するアルゴリズム) によって視覚化している．このアルゴリズムだと，関連性の高いノード (名詞) が近くに集まるはずである．

確かに，一部のノード (例えば右下の「日本国」,「統合」,「総意」など) が独立した集団 (コミュニティ) を形成しいるが，大半のノードは中央の巨大なコミュニティの中に埋もれてしまい，この図を用いて「概念の理解を手助けする」ということは難しいと考えられる．

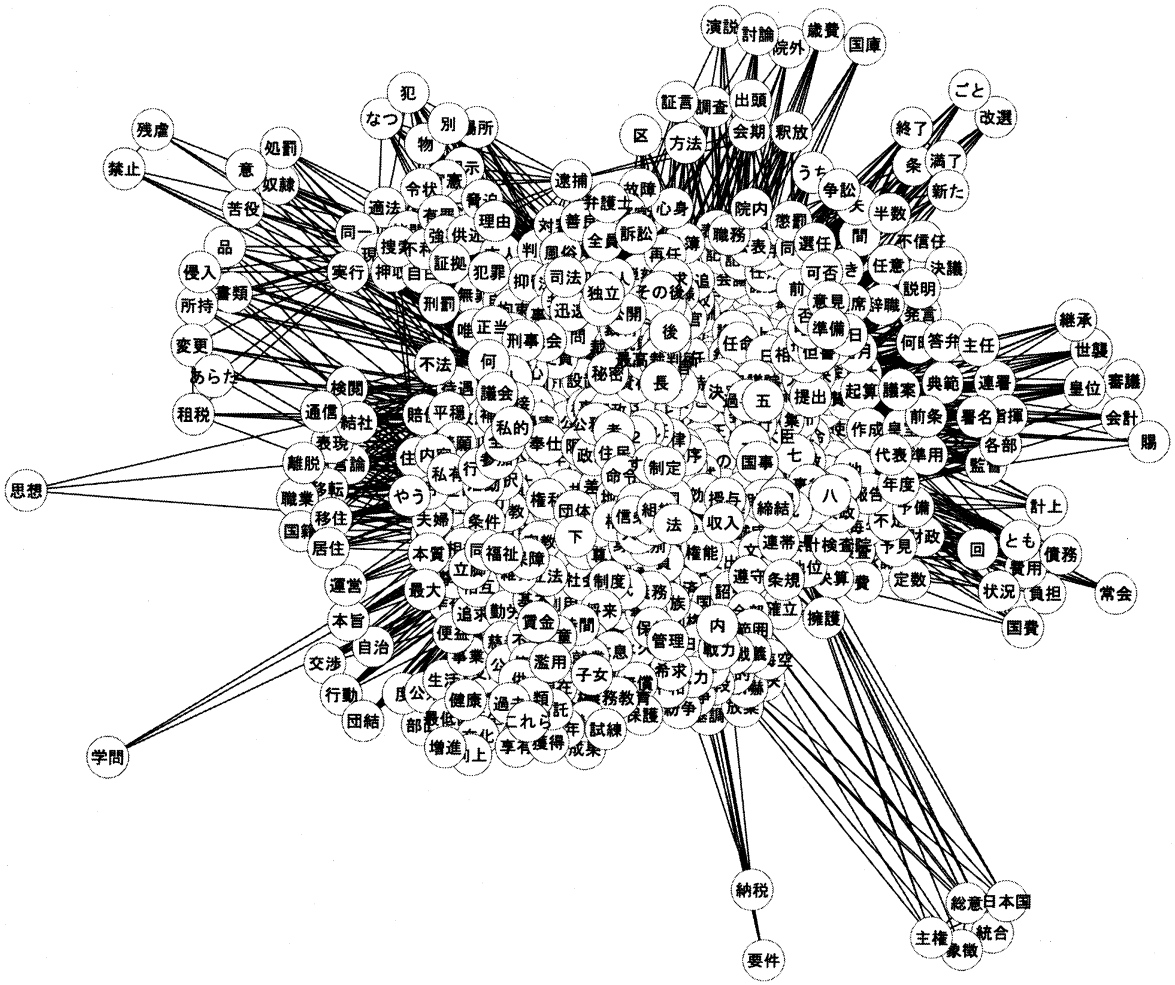

図 2: 日本国憲法全体のネットワーク

2.3 一般性の高い名詞の除去によるノードの刈り込み

図2には，日本国憲法に出現するすべての名詞が含まれている．そのために，図全体が煩雑なものとなってしまっている．特に，図2を見ると,「何」,「後」,「下」といった，それ自身が法律用語として重要な意味を持つとは考えにくい単語が散見される．

ところが，こういった一般的な名詞は，法文の随所で用いられるため,「概念の理解の手助け」という観点からは，あまり意味のないネットワーク構造を形成してしまっていると考えられる．

そこで，こういったネットワークの構造において重要性が薄いと考えられる名詞(ノード)の数を減らす事をまず試みた．具体的には，次のような名詞を除去の対象とした．なお，重要性の高低に関しては，現時点では筆者らの判断によっている．

- 1，2，3，一，二，三，といった数詞
- これ，それ，といった指示代名詞
- その後，こと，すべて，場合，的，といった法的に意味の薄いと考えられる名詞

- 形態素解析の誤認識によるとみられる名詞

図3は，以上の処理を施した状態である．この状態で，ネットワークは，当初の状態から74個のノードが除去され，523個のノードと，8813本のエッジから構成されている．

残念ながら，この状態でも，ほぼすべてのノードが一つのコミュニティに集まってしまい，何かの構造を示唆するような視覚化の効果は得られていない．

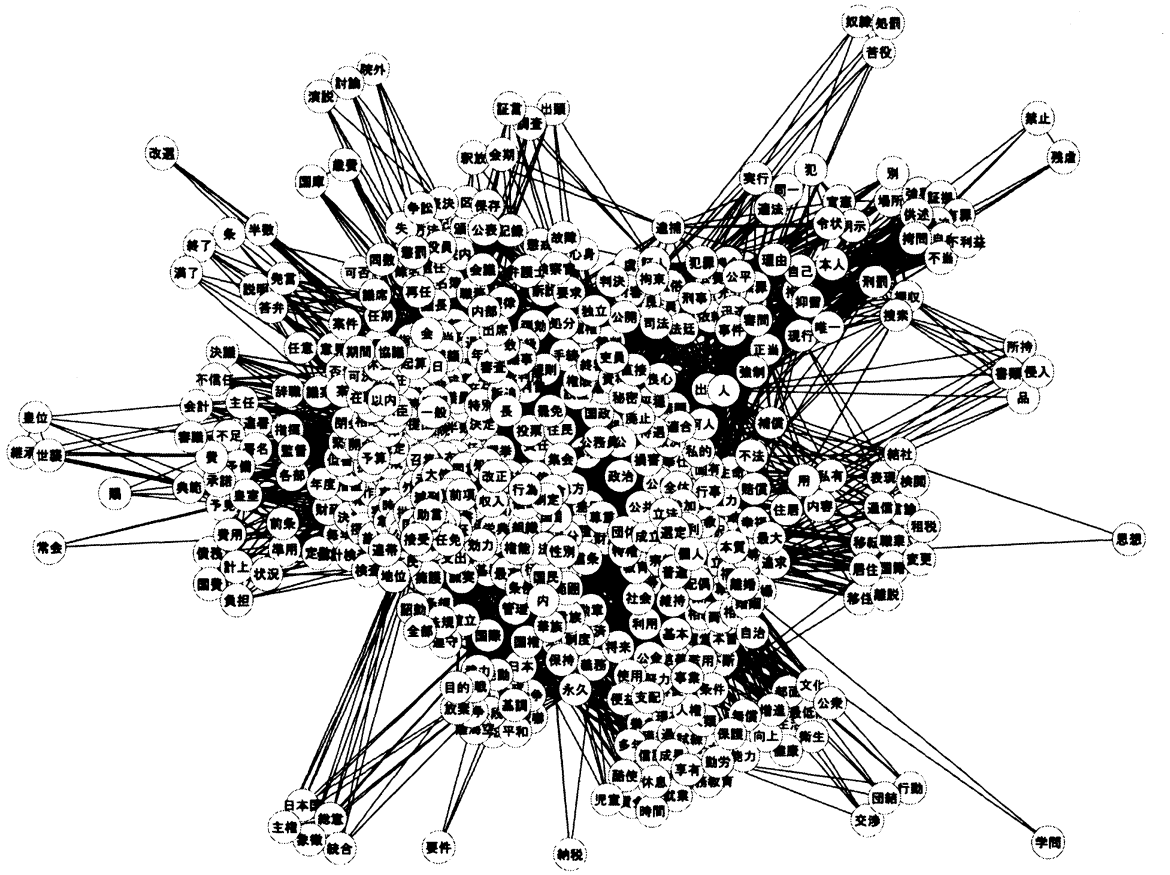

図 3: 一般性の高い名詞を除去した状態

2.4 中心性によるノードの刈り込み

ノードの中心性 [1, 2, 3, 4] とは，直観的には，ネットワークにおける当該ノードの重要性を意味する．次数中心性，近接中心性，媒介中心性といった各種の中心性が定義されている．

そこで，この中心性の高いノードだけを抽出し，それらを視覚化する事で，何か意味のあるネットワーク構造が見えて来るのではないかということが期待できる．

本研究では，Cytoscape のプラグインの一つである cytoHubba[11] を用いて，前節で述べた一般性の高い名詞除去後のネットワークに対して，複数の種類の中心性について視覚化を試みた．

一般的な中心性の指標である，エッジの次数中心性，媒介中心性などでは，必ずしも良好な視覚化の結果が得られなかった．

その中で，Chung-Yen Lin らによって提案されたDMNC[5] と呼ばれる指標を利用した視覚化が，きわめて明瞭なネットワーク構造を示すことが判明した．

いま，ネットワークが無向グラフ $G = (V, E)$ で与えられるとする．ここで，V はノード (Vertex)，E はエッジの集合である．

この時，DMNC(Density of Maximum Neighborhood Component) は，次のように定義される．

$$DMNC(v) = \frac{|E(MNC(v))|}{|V(MNC(v))|^{\varepsilon}} \tag{1}$$

ここで，$MNC(v)$ は，あるノード $v(\in V)$ に隣接するノードの集合 (v 自身は除く) を $N(v)$ とし，$N(v)$ によって構成される部分グラフの「最大の連結成分」(maximum connected component) である．[1]

なお，$|E(G)|$，$|V(G)|$ は，グラフ G のエッジ数、ノード数である。cytoHubba では，ε は 1.7 に設定されている．

図 4 は，DMNC 上位 100 ノードによる視覚化である．この図に示すように，ネットワークは複数の独立したコミュニティに分離する．出現する名詞から判断して，例えば上段に関しては，左から，それぞれ下のような条文に相当するものと考えられる．

- 第 7 条「天皇の国事行為」および第 73 条「内閣の職務」
- 第 9 条「戦争の放棄，軍備及び交戦権の否認」
- 第 24 条「家庭生活における個人の尊厳と両性の平等」
- 第 79 条「最高裁判所の裁判官国民審査，定年，報酬」および 第 80 条「下級裁判所の裁判官・任期定年，報酬」

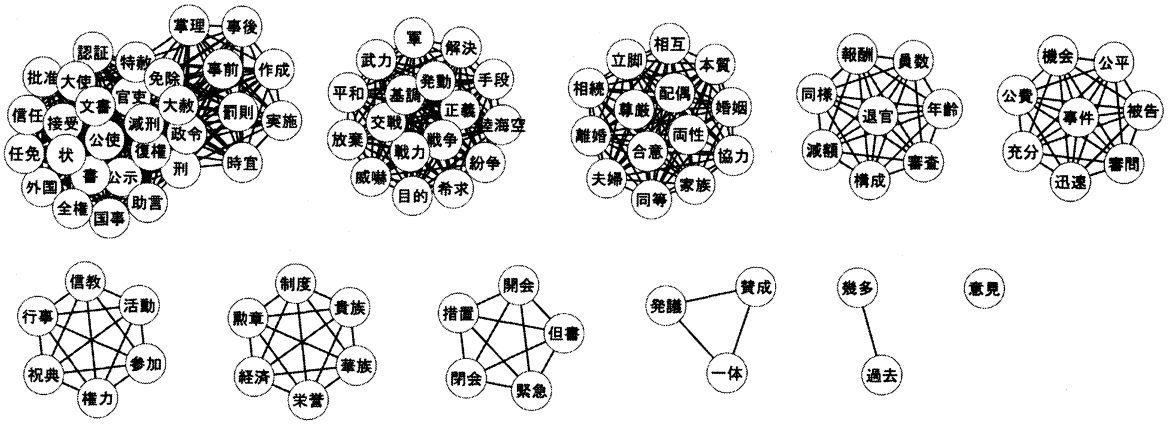

図 4: DMNC 上位 100 ノードによる視覚化

[1]直観的な言い方をすれば，自分の直接の友人グループの中の最大派閥，と考えればよい．

- 第37条「刑事被告人の権利」

ここで注意すべきは，複数の条文が一つのコミュニティを形成する場合がある，という点である．

例えば，第7条「天皇の国事行為」および第73条「内閣の職務」は，条文を読めば確かに類似点が多いことは明らかであるが，機械的な手法でその類似性が抽出されているということになる．

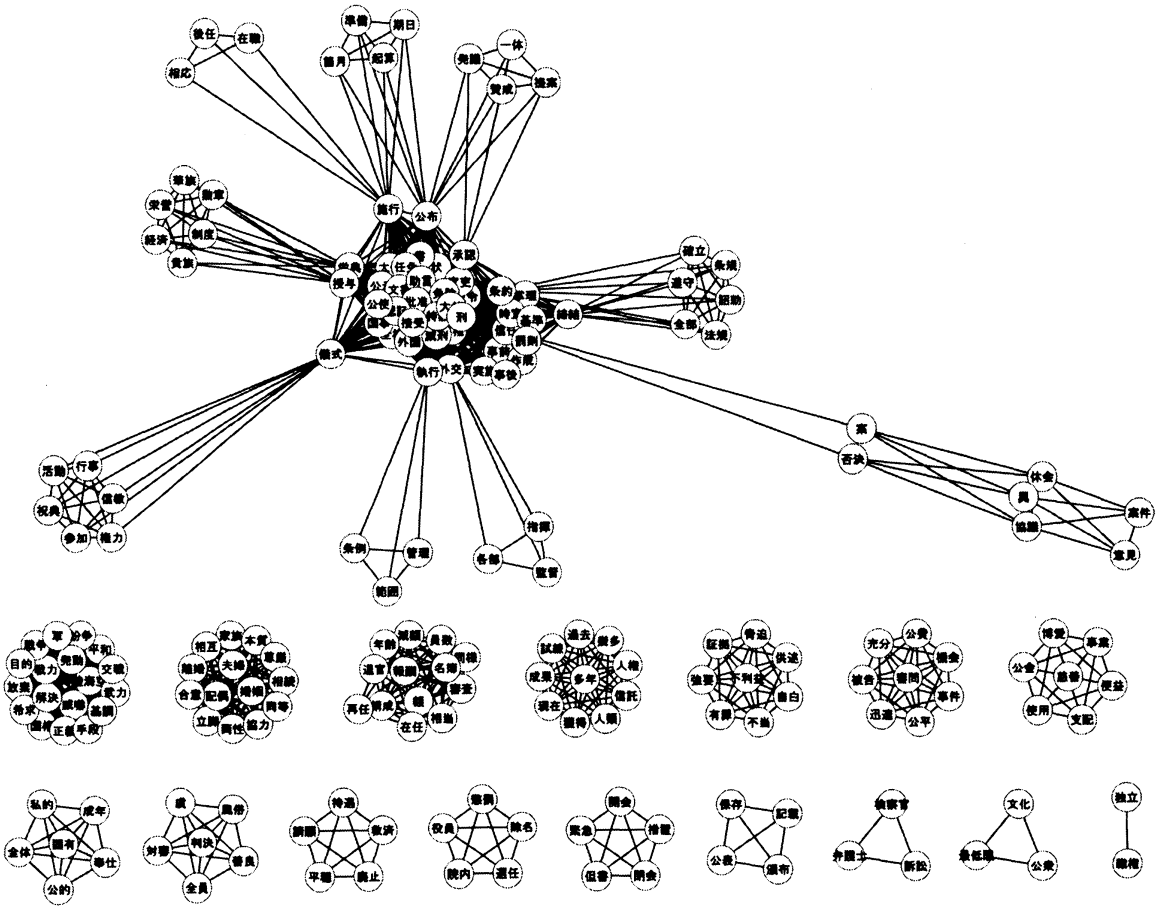

図 5: DMNC 上位 200 ノードによる視覚化

図5は，DMNC上位200ノードによる視覚化である．この場合は，おおむね一つないし少数の条文に対応する小さなコミュニティが図の下部に並ぶ一方で，図の上部に複数のコミュニティが連結した大きなネットワークが現れる．

この時，コミュニティ同士が少数の名詞を介して連結している場合が見られる．

例えば，この大きなネットワークの左下の「行事」,「信教」といった名詞を含むコミュニティは，憲法第20条「信教の自由」に相当すると考えられるが，このコミュニティは「儀式」という名詞を媒介して, ネットワークの中央部分と連結している．

したがって,「儀式」という名詞は，条文間の関連を示す重要な単語である，という可能性が指摘できる．

このように，DMNCという中心性を表す指標を用いて，100ノードあるいは200ノードといったサブネットワークを抽出することで，明瞭なコミュニティ構造を見出すことが可能となってくる．

このことは，やはり，ネットワークの可視化において，何らかの指標による「刈り込み」の重要性を示しているものと考えられる．

2.5 DMNCはなぜ有効なのか

ではなぜ、DMNCが他の中心性に比べて視覚化に有効なのであろうか?

DMNCは，あるノードvに隣接するノードの集合$N(v)$を用いて定義されている．一方，視覚化の対象としたネットワークは，一つの条文の中に出現する名詞には「相互に関係がある」と考えて作ったものである．したがって，一つの条文だけを考えれば，条文内のある名詞vに対する$N(v)$は，条文内の他の名詞で構成される完全グラフになっている．

それ故，比較的長い条文で，その中に現れる各名詞のDMNCの値が高くなるのは当然の事である．また，複数の条文で共通する名詞が，それぞれの条文のコミュニティを連結する位置に来ることも，当然である．

このように，DMNC上位ノードを用いた視覚化では，法律にもともとあった条文という単位に強く依存するノードの刈り込みが行われていたことになる．

端的に言えば，DMNCは,「もともとの文章が明瞭な単位に区分されている場合に有効な指標」である，と考えることができる．逆に，そのような区分がない場合は，その有効性は必ずしも明らかではない．

3 おわりに

本研究では,「複雑ネットワークの知見やツールを利用して，複雑な概念を(可能な限り)機械的に視覚化し，概念の理解を手助けすることが可能かどうか，検証を試みる」という方針に基づき，法律(日本国憲法)の条文に出現する名詞によるネットワークを構成し，その視覚化を行った．

その結果，少なくとも対象とした日本国憲法に関しては，単純な視覚化では意味のある構造を見出すことは困難であった．

しかし，DMNCという中心性に関する指標を用い，その上位ノードだけを選択的に視覚化することで，条文に対応するコミニュティと，それらのコミュニティが相互に連結している構造が視覚化できることが判明した．

これは，当初の目標である「複雑な概念を(可能な限り)機械的に視覚化し，概念の理解を手助けすること」が，条件次第である程度は実施可能であることを示唆する．

今回は，視覚化ツール(Cytoscape)にあらかじめ用意されている機能のみを用いたが，適切な前処理や表示の工夫を行えば，より直観的に理解しやすい構造の表示が実施できるのではないかと期待できる．

また，日本国憲法以外の法律，あるいは，法律以外の文章の場合は，どのような傾向を示すのか検証が必要である．

さらに，DMNC 以外にも，視覚化のための有用な指標が存在するかどうかの検討も将来的な課題である．

参考文献

[1] 安田雪：実践ネットワーク分析, 新曜社, 2001.

[2] 増田直樹，今野紀雄：複雑ネットワークの科学，産業図書, 2005.

[3] 林幸雄編著，大久保潤，藤原義久，上林憲行，小野直亮，湯田聰夫，相馬亘，佐藤一憲：ネットワーク科学の道具箱，近代科学社, 2007.

[4] 今野紀雄，町田拓也：図解入門よくわかる複雑ネットワーク，秀和システム, 2008.

[5] Lin et al : Hubba: hub objects analyzer - a framework of interactome hubs identification for network biology, Oxford Journals, Nucleic Acids Research, Volume 36 suppl 2, 2008, http://nar.oxfordjournals.org/content/36/suppl_2/W438.full

[6] 金明哲：テキストデータの統計科学入門，岩波書店, 2009.

[7] 吉澤康介, 三宅修平：インターネット通販サイトにおける関連商品情報等の抽出と可視化の試み, 第 7 回ネットワーク生態学シンポジウム予稿集, 情報処理学会研究報告, 2011.

[8] http://www.cytoscape.org/

[9] http://pajek.imfm.si/doku.php

[10] http://med.bioinf.mpi-inf.mpg.de/netanalyzer/

[11] http://hub.iis.sinica.edu.tw/cytoHubba/

[12] http://chasen-legacy.sourceforge.jp/